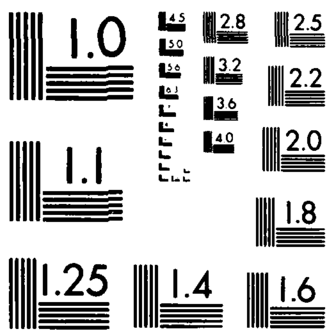
AD-A149 407   ESTIMATION OF VARIANCE OF THE REGRESSION ESTIMATOR(U)   1/1
WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH CENTER
L Y DENG ET AL. OCT 84 MRC-TSR-2758 DAAG29-80-C-0041
UNCLASSIFIED   F/G 12/1   NL

END
FILMED
DTIC



MICROCOPY RESOLUTION TEST CHART

NATIONAL BUREAU OF STANDARDS-1963-A

3

AD-A149 407

MRC Technical Summary Report #2758

ESTIMATION OF VARIANCE OF THE REGRESSION ESTIMATOR

Lih-Yuan Deng and C. F. Jeff Wu

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53705

October 1984

(Received September 11, 1984)

Approved for public release
Distribution unlimited

Sponsored by

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

National Science Foundation
Washington, DC 20550

UNIVERSITY OF WISCONSIN - MADISON
MATHEMATICS RESEARCH CENTER

ESTIMATION OF VARIANCE OF THE REGRESSION ESTIMATOR

Lih-Yuan Deng* and C. F. Jeff Wu**

Technical Summary Report #2758

October 1984

ABSTRACT

For estimating the variance of the regression estimator in simple random sampling without replacement, several design-based and model-based estimators and a new class of estimators are compared. Their second order expressions and biases are derived and compared. Empirical results on the biases and MSE's (Mean Squared Errors) of the variance estimators and the conditional and unconditional coverage probabilities of their associated t-intervals lend support to the theoretical results and suggest further questions. Originator-supplied

AMS (MOS) Subject Classification: 62D05

Key Words: Variance estimator; Design-based; Model-based; ~~Jackknife;~~ Conditional coverage probabilities. Jackknife

Work Unit Number 4 - Statistics and Probability

*Assistant Professor, Department of Mathematical Sciences, Memphis State University, Memphis, TN 38152.

**Professor, Department of Statistics, 1210 W. Dayton St., University of Wisconsin-Madison, Madison, WI 53706.

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041. This research is supported by the Graduate School Research Committee, University of Wisconsin-Madison and the National Science Foundation under Grant No. MCS-8300140.

SIGNIFICANCE AND EXPLANATION

In estimating the population mean of a character y, we often make use of an auxiliary covariate x about which information is more readily available and is positively correlated with y. One commonly used estimator in survey sampling is the regression estimator. To assess the variability of the estimator, we need an estimator for its variance. Several variance estimators have been proposed using model-based or design-based arguments. We propose a class of variance estimators, which includes or approximates several existing variance estimators in the literature. The asymptotic variance and bias of these estimators are found and compared with results from an empirical study. Empirical results on coverage probabilities of Student's t-intervals with these variance estimators are also obtained and proper interpretation is given.

The responsibility for the wording and views expressed in this descriptive summary lies with MRC, and not with the authors of this report.

ESTIMATION OF VARIANCE OF THE REGRESSION ESTIMATOR

Lih-Yuan Deng* and C. F. Jeff Wu**

1. Introduction

The main purpose of this paper is to provide a theoretical and empirical comparison of several variance estimators for the regression estimator in simple random sampling without replacement. The companion problem for the ratio estimator has been well studied in the literature. See the references of Wu and Deng(1983) and Rao(1985). In the past more attention has been given to the ratio estimator because of its computational ease and general applicability for general sampling designs. The ratio estimator is appropriate for populations whose regression line passes close to the origin. If the intercept of the regression line is significantly nonzero, it is much less efficient than the regression estimator(Deng, 1984). In general, apart from n^{-2} terms, the mean squared error of the former is bigger than that of the latter(Cochran, 1977, p.196). For estimating cell totals in tables of the type typically constructed from survey data, Fuller(1977) showed the superior performance of the regression estimator. For stratified samples Wu(1985) showed that the model underlying the use of the combined ratio estimator has an artificial constraint while the model for the combined regression estimator is more natural. Given the present availability of fast and inexpensive computing, the computational advantage of the ratio estimator should be less of a concern and the regression estimator will gain wider popularity.

*Assistant Professor, Department of Mathematical Sciences, Memphis State University, Memphis, TN 38152.

**Professor, Department of Statistics, 1210 W. Dayton St., University of Wisconsin-Madison, Madison, WI 53706.

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041. This research is supported by the Graduate School Research Committee, University of Wisconsin-Madison and the National Science Foundation under Grant No. MCS-8300140.

There are two approaches to the variance estimation problem. The traditional one, based on the probability distribution generated by the sampling design, is well summarized in Cochran's book. By imposing a superpopulation model on the actual finite population, inference about the characteristics of the finite population can be made via the structure of the model(Brewer, 1963; Scott and Smith, 1969; Royall, 1970). Several model-based variance estimators were proposed and studied in Royall and Eberhardt(1975), Royall and Cumberland(1978). For the regression estimator, an empirical study of these model-based variance estimators and a traditional estimator v_{1r} (2.7) was given in Royall and Cumberland(1981). Several traditional estimators were compared in earlier studies by Rao(1968, 1969). The estimators in the above three papers and some new ones(formula (2.9)) will be studied in our paper. Our theoretical comparison of these design-based and model-based variance estimators is design-based, although some results are given a model-based interpretation. More precise results are made possible by the second order expansions of these estimators reported in Section 3. Our simulation study contains two new features, the mean squared errors (MSE) of the variance estimators and the conditional coverage probabilities of the associated t-intervals.

The organization and major findings of this paper are as follows. Section 2 lists all the variance estimators under comparison, including a class of adjustments, (2.9), of the standard variance estimator v_{1r} (2.7). The optimal adjustment within the class (2.9) is studied in Section 3.2 in parallell to Wu(1982a). From their respective asymptotic expansions, the jackknife estimator v_J (2.20) and two bias-robust estimators v_D (2.13) and v_H (2.14) have the same leading term of order n^{-1}. For the next order terms, v_J is bigger than v_D, which in turn is bigger than v_H. The same expansions also enable us

to compute the biases of these estimators for estimating the MSE of the regression estimator $\hat{\bar{y}}_{1r}$ (2.1). To achieve this goal, a new expansion (to order n^{-2}) for MSE($\hat{\bar{y}}_{1r}$) is derived in Theorem 4.1. Among all the estimators, v_D is the only one that captures the n^{-1} and n^{-2} terms of MSE($\hat{\bar{y}}_{1r}$). Its absolute bias is of order $n^{-2.5}$ and is the smallest. The jackknife estimator v_J overestimates MSE($\hat{\bar{y}}_{1r}$) while v_H underestimates MSE($\hat{\bar{y}}_{1r}$). A condition (4.7) (which is often satisfied by natural populations) is found, under which the commonly used estimator v_{1r} and another one v_L underestimate MSE($\hat{\bar{y}}_{1r}$). The findings on bias are well supported by Royall-Cumberland's(1981) study (summarized in Table 1) and our study in Section 5(Table 2). The empirical MSE behavior (Table 2) of different variance estimators support the theoretical result Theorem 3.1. Those v_g with g chosen to be g_{opt} (2.10) have smaller MSE's. An interesting and somewhat surprising finding is that the jackknife variance estimator v_J consistently has the largest MSE. Typically the two model-based estimators v_D and v_H have bigger MSE's. If the MSE of $\hat{\bar{y}}_{1r}$ is the primary parameter of interest as in determining the sample size for future surveys, the optimal estimator $v_{\hat{g}}$ should be used in place of v_J, v_H or v_D. For coverage probabilities of t-intervals of the form (5.2), which are relevant to internal inference about the population mean, we observe a reverse pattern. In terms of the closeness of the empirical unconditional coverage probabilities to the nominal level (Table 3), we have

$v_J > v_D > v_H > v_2 > v_1 > v_{1r}$ in decreasing order of performance. In terms of the stability and closeness (to the nominal level) of the coverage probabilities conditional on the sample mean of the covariate, a similar pattern is observed. This is interesting since the losers v_J, v_D and v_H for estimating MSE($\hat{\bar{y}}_{1r}$) turn out to be the big winners here. Perhaps the most important recommendation for practitioners is that the commonly used estimator v_{1r} fails on both grounds and should only be used with caution. An obvious conclusion is that different variance estimators should be used for different purposes. Further theoretical study is needed to understand this empirical phenomenon (the same phenomenon was observed in Wu and Deng's empirical study for the ratio estimator.)

The restriction to simple random sampling without replacement will undoubtedly rule out many large scale complex surveys. We hope our study will inspire further interest and eventually lead to useful recommendations for more complex situations. In settings like marketing research, simulation analysis (Iglehart, 1978) and telephone surveys where simple random sampling is a key element of the sampling plan, our results may be directly applicable.

2. Variance Estimation For Regression Estimator

Consider a population consisting of N distinct units with values (x_i, y_i) , i=1(1)N , with x_i positive and known. Samples are drawn from the population at random without replacement. Denote the sample and population means of y_i and x_i by $\bar{y}$, $\bar{x}$ and $\bar{Y}$, $\bar{X}$ respectively.

Two estimators of $\bar{Y}$ commonly used in practice are the ratio estimator

$$\hat{\bar{y}}_R = \frac{\bar{X}}{\bar{x}}\,\bar{y}$$

and the regression estimator

$$\hat{\bar{y}}_{lr} = \bar{y} + b(\bar{X} - \bar{x}) , \tag{2.1}$$

where

$$b = \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) / \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{2.2}$$

is the sample regression coefficient of y_i on x_i. The regression estimator is the best linear unbiased predictor of $\bar{Y}$ under the following superpopulation model (Royall, 1970)

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{2.3}$$

where ε_i are uncorrelated with mean zero and variance σ^2. The superpopulation model underlying the use of the ratio estimator is the one without the intercept term β_0.

The leading term of the mean squared error (MSE) or variance of $\hat{\bar{y}}_{lr}$ is

$$V = \left(\frac{1-f}{n}\right)\frac{1}{N-1} \sum_{i=1}^{N} e_i^2 , \tag{2.4}$$

where

$$e_i = (y_i - \bar{Y}) - B(x_i - \bar{X}) \tag{2.5}$$

is the <u>residual</u> of y_i to the regression line $\bar{Y} + B(x_i - \bar{X})$,

$$B = \sum_{i=1}^{N} (x_i - \bar{X})(y_i - \bar{Y}) / \sum_{i=1}^{N} (x_i - \bar{X})^2 \tag{2.6}$$

is the population regression coefficient of y_i on x_i, and $f = n/N$ is the sampling fraction.

The most commonly used estimator of the approximate variance V is its sample analogue

$$v_{1r} = \left(\frac{1-f}{n}\right)\frac{1}{n-2}\sum_{i=1}^{n}\hat{e}_i^{\,2} , \tag{2.7}$$

where

$$\hat{e}_i = (y_i - \bar{y}) - b(x_i - \bar{x}) \tag{2.8}$$

is the i-th residual based on the sample and b is given in (2.2).

For estimating the variance of the ratio estimator, Wu(1982a) considered $v_g = \left(\frac{\bar{X}}{\bar{x}} \right)^g v_0$ as a class of adjustments of the usual estimator (Cochran, 1977, p.155)

$$v_0 = \left(\frac{1-f}{n}\right)\frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \frac{\bar{y}}{\bar{x}} x_i\right)^2 . \tag{2.8.1}$$

He then proposed to choose g by minimizing the mean squared error of v_g. In an empirical study by Wu and Deng (1983), the optimal v_g performs well among several other variance estimators. In the regression case we will consider a similar class of variance estimators

$$v_g = \left(\frac{\bar{X}}{\bar{x}} \right)^g v_{1r} . \tag{2.9}$$

Let S_{zx} denote the population covariance of x_i and z_i, S_x^2 the population variance of x_i. It will be shown in Theorem 2.1 that the leading terms of MSE(v_g) is minimized by

$$g_{opt} = \frac{S_{zx}/ \bar{X}\,\bar{Z}}{S_x^2/ \bar{X}^2} \tag{2.10}$$

which is the population regression coefficient of $z_i/ \bar{Z}$ over $x_i/ \bar{X}$, i= 1(1)N and $z_i = e_i^2$ is the residual squared. This suggests the following optimal estimator within the class (2.9),

$$v_{\hat{g}} = \left(\frac{\bar{X}}{\bar{x}}\right)^{\hat{g}} v_{1r} , \qquad (2.11)$$

where $\hat{g}$ is the sample analog of g_{opt}.

For variance estimation of the ratio estimator, Fuller(1981) suggested a regression adjustment to v_0 (2.8.1). A similar adjustment can be applied to v_{1r}. For the ratio estimator, as pointed out in Wu and Deng (1983), Fuller's estimator is asymptotically equivalent to $(\bar{X}/\bar{x})^{\hat{g}} v_0$, where $\hat{g}$ is the sample analogue of the optimal g_{opt}. The corresponding result is also true for the regression estimator.

Another variance estimator closely related to v_{1r} is

$$v_L = v_{1r}\left[1 + \frac{(\bar{x} - \bar{X})^2}{(\frac{1-f}{n}) \sum_{i=1}^{n} (x_i - \bar{x})^2}\right] , \qquad (2.12)$$

whose justification comes from standard regression theory (Cochran, 1977 , p.199).

Royall and Cumberland (1978) proposed two bias-robust (against misspecification in model (2.3)) variance estimators

$$v_D = \frac{(1-f)^2}{n(n-1)} \sum_{i=1}^{n} \alpha_i \hat{e}_i^2 \qquad (2.13)$$

$$v_H = \left(\frac{1-f}{n}\right)^2 \sum_{i=1}^{n} \beta_i \hat{e}_i^2 + f\left(\frac{1-f}{n}\right)\frac{1}{n-2} \sum_{i=1}^{n} \hat{e}_i^2 , \qquad (2.14)$$

where

$$\alpha_i = \frac{r_i^2 + f/(1-f)}{1-(x_i - \bar{x})^2/((n-1)g(s))} , \qquad (2.15)$$

$$g(s) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 , \quad \bar{x}_r = \frac{N\bar{X} - n\bar{x}}{N-n} , \qquad (2.16)$$

$$r_i = 1 + (x_i - \bar{x})(\bar{x}_r - \bar{x})/g(s) \tag{2.17}$$

and

$$\beta_i = r_i^2/(1-\frac{1}{n}\sum_{i=1}^{n} w_i k_i), \quad w_i = r_i^2/\sum_{i=1}^{n} r_i^2, \tag{2.18}$$

$$k_i = [1 + (x_i - \bar{x})^2/g(s)]/n . \tag{2.19}$$

The last estimator under comparison is the jackknife variance estimator

$$v_J = (\frac{1-f}{n})(n-1)\sum_{i=1}^{n}(\hat{\theta}_{(i)}-\hat{\theta}_{(.)})^2 , \tag{2.20}$$

where $\hat{\theta}_{(i)}$ is the regression estimate (2.1) based on the sa e of size n-1 with unit i deleted from the sample and $\hat{\theta}_{(.)}$ is the ge of $\hat{\theta}_{(i)}$.

3. Relationships among the Variance Estimators under Comparison

3.1. Asymptotic Expansions

To study the asymptotic relationships among the variance estimators in Section 2, we need the following asymptotic expansions

$$\delta_n = b-B = \sum_{i=1}^{n}(x_i - \bar{x})(e_i - \bar{e})/\sum_{i=1}^{n}(x_i - \bar{x})^2 = O_p(n^{-0.5}) \tag{3.1}$$

$$= [\bar{u} - \bar{e}(\bar{x} - \bar{X})]/(n^{-1}(n-1)s_x^2) = \bar{u}/S_x^2 + O_p(n^{-1}) \tag{3.2}$$

$$= \frac{\bar{u} - \bar{e}(\bar{x} - \bar{X})}{S_x^2} - \frac{\bar{u}(\bar{v} - \bar{V})}{S_x^4} + O_p(n^{-1.5}) , \tag{3.3}$$

where

$$u_i = e_i(x_i - \bar{X}) , \quad v_i = (x_i - \bar{X})^2 \tag{3.4}$$

and $\bar{u}, \bar{v}$ are the sample means of u_i and v_i, $\bar{V}$ the population mean of v_i. Since the population means of e_i and u_i are zero,

$$\bar{e} = O_p(n^{-0.5}) , \quad \bar{u} = O_p(n^{-0.5}) . \tag{3.5}$$

Ignoring the lower order terms of δ_n^2, we have

$$\delta_n^2 = O_p(n^{-1}) = \frac{\bar{u}^2}{S_x^4} + O_p(n^{-1.5}) \tag{3.6}$$

$$= \frac{\bar{u}^2 - 2\ \bar{u}\ \bar{e}(\ \bar{x} - \bar{X})}{S_x^4} - 2\frac{\bar{u}^2(\bar{v} - \bar{V})}{S_x^6} + O_p(n^{-2}) . \tag{3.7}$$

In writing (3.3) and (3.7), we used $s_x^2 = S_x^2 + O_p(n^{-0.5})$.

3.2. Optimal Variance Estimators among v_g

Using the minimum mean squared error of the variance estimator as the criterion, we will choose an optimal estimator within the class (2.9). The following lemma finds the leading terms of v_g and Var(v_g)

Lemma 3.1.

(a) $v_{1r} = (\frac{1-f}{n})\ \bar{z} + O_p(n^{-2})$, where $\bar{z} = \frac{1}{n} \sum_{i=1}^{n} z_i$, $z_i = e_i^2$.

(b) $v_g = (\frac{1-f}{n})(\ \bar{z} + g(\delta\ \bar{x})\ \bar{z}) + O_p(n^{-2})$, where $\delta\ \bar{x} = \frac{(\ \bar{x} - \bar{X})}{\bar{X}}$.

(c) $\text{Var}(\ v_g) = (\frac{1-f}{n})^3 (\ S_z^2 - 2\ g\ (\frac{\bar{Z}}{\bar{X}})\ S_{zx} + g^2\ (\frac{\bar{Z}}{\bar{X}})^2\ S_x^2) + O(n^{-3.5})$.

Except for the obvious ones, the derivations and proofs in this paper are given in the Appendix.

By minimizing expression (c) of Lemma 3.1, we have

Theorem 3.1. The optimal choice of g, minimizing the variance of v_g, is given by g_{opt} defined in (2.10).

For estimating the variance of the ratio estimator, a similar result to Theorem 3.1 was obtained in Wu(1982a) with a major difference. His z_i takes a more complex form

$$d_i^{\,2} - 2\,\frac{\sum_{i=1}^{N} x_i d_i}{\sum_{i=1}^{N} x_i}\, d_i \quad , \tag{3.8}$$

where $d_i = y_i - (\bar{Y}/\bar{X})\, x_i$ is the residual in the ratio context. Note that the second term of (3.8) does not appear in the regression case. One explanation for this difference is that the regression estimator incorporates a non-zero intercept term while the ratio estimator suppresses it. More precisely, each y value can be decomposed as

$$y_i = A + B\, x_i + e_i \tag{3.9}$$

where B and e_i are defined in (2.6) and (2.5), $A = \bar{Y} - B\bar{X}$ is the intercept from fitting a regression line to the population $(y_i, x_i), i=1(1)N$. With this representation, $d_i = -A(x_i - \bar{X})/\bar{X} + e_i$ and

$$\sum_{i=1}^{N} x_i\, d_i = -A\,\frac{(N-1)\, S_x^2}{\bar{X}} \quad , \tag{3.10}$$

from which it is easy to see that the extra term in (3.8) would be zero if the intercept were zero.

To obtain further properties of v_g, let us assume the superpopulation model

$$y_i = \alpha + \beta\, x_i + \varepsilon_i \tag{3.11}$$

$$E_M(\varepsilon_i) = 0 \;;\; E_M(\varepsilon_i\, \varepsilon_j) = \sigma^2 x_i^{\,t} \text{ for } i=j;\ 0 \text{ for } i \neq j,$$

where E_M denotes expectation with respect to the model. Under (3.11), $B = \beta + O(N^{-1/2})$ and $e_i = \varepsilon_i + O(N^{-1/2})$. By using Wu's(1982a) argument, we find that up to order N^{-1},

$$g_* = \frac{\left(\sum_{i=1}^{N} x_i^{t+1} - \bar{X} \sum_{i=1}^{N} x_i^{t} \right) \left(\sum_{i=1}^{N} x_i \right)}{\sum_{i=1}^{N} (x_i - \bar{X})^2 \sum_{i=1}^{N} x_i^{t}} ,$$

minimizes E_M (Var(v_g)) under (3.11), from which the following results are readily obtained. Compare with Wu(1982a, Propositions 1 and 2).

Theorem 3.2. Under model (3.11) with

(a) t=0, v_0(= v_{1r}) is the optimal estimator of V among v_g;

(b) t=1, v_1 is the optimal estimator of V among v_g;

(c) $t \geq 1$, then $g_* \geq 1$ and v_1, v_2 are both better than v_0 for estimating V.

Recall that under (3.11) with t=0 , $\hat{\bar{y}}_{1r}$ is the best linear unbiased predictor of $\bar{Y}$.

3.3. Relationships among v_D, v_H and v_J

The two estimators v_H and v_D are approximately unbiased estimators of the true error variance even when the error variance structure is not correctly specified by the model. According to Theorem 3 of Royall and Cumberland(1978), under some mild conditions, v_H , v_D and v_J are asymptotically equivalent, i.e., $v_H = v_J(1 + o(1))$ and so on. By studying the second order terms of the variance estimators, we find some interesting relationships among them. We will show that v_J is stochastically larger than v_D and v_D is larger than v_H. Lemmas 3.2 and 3.3 find the leading terms of v_D and v_H. Throughout this subsection, we assume f = $O(n^{-0.5})$.

Lemma 3.2.

$$v_D = \frac{1-f}{n(n-1)} \sum_{i=1}^{n} \hat{e}_i^2 \frac{(1-p(x_i-\bar{x}))^2}{1-q(x_i-\bar{x})^2} + O_p(n^{-2.5}) \tag{3.12}$$

$$= \frac{1-f}{n(n-1)} \sum_{i=1}^{n} \hat{e}_i^2 [1-2p(x_i-\bar{x})+(p^2+q)(x_i-\bar{x})^2] + O_p(n^{-2.5}) \tag{3.13}$$

where

$$p = \frac{(\bar{x}-\bar{X})}{g(s)}, \quad q = \frac{1}{(n-1)g(s)} \tag{3.14}$$

and $g(s)$ is defined in (2.16).

Lemma 3.3.

$$v_H = \frac{1-f}{n^2} \sum_{i=1}^{n} (1-p(x_i-\bar{x}))^2 \hat{e}_i^2 + O_p(n^{-2.5}) . \tag{3.15}$$

From (3.13) and (3.15), we have

Lemma 3.4.

$$v_H - v_D = -\frac{1-f}{n(n-1)}\Big[\frac{1}{n} \sum_{i=1}^{n} \hat{e}_i^2 (1-p(x_i-\bar{x}))^2$$

$$+ q \sum_{i=1}^{n} \hat{e}_i^2 (x_i-\bar{x})^2\Big] + O_p(n^{-2.5}) . \tag{3.16}$$

Lemma 3.4 implies that v_D is asymptotically larger than v_H. Lemma 3.5 finds the leading terms of v_J.

Lemma 3.5.

$$v_J = \frac{1-f}{n(n-1)} \sum_{i=1}^{n} \frac{\hat{e}_i^2 (1-p(x_i-\bar{x}))^2}{(1-q(x_i-\bar{x})^2)^2} + O_p(n^{-2.5}) . \tag{3.17}$$

We can compare v_D and v_J based on Lemmas 3.2 and 3.5.

Lemma 3.6.

$$v_J = v_D + q \frac{1-f}{n(n-1)} \sum_{i=1}^{n} \hat{e}_i^2 (x_i - \bar{x})^2 + O_p(n^{-2.5}) . \tag{3.18}$$

Lemma 3.6 implies that v_J is asymptotically larger than v_D.

4. Asymptotic Bias Behavior of Variance Estimators

4.1. Second-order Expansions of MSE($\hat{\bar{y}}_{1r}$) and v_{1r}

Theorem 4.1. Let V be the approximate variance (2.4).

(a) $\text{MSE}(\hat{\bar{y}}_{1r}) = V$

$$+\left(\frac{1-f}{n}\right)^2 \left(2 S_e^2 - \frac{1-2f}{1-f} \frac{S_u^2}{S_x^2} + \frac{4 S_{x^2e}^2 + 2 S_{xe^2} U_3}{S_x^4}\right) + O(n^{-2.5}),$$

where $U_3 = (N-1)^{-1} \sum_{i=1}^{N} (x_i - \bar{X})^3$, S_u^2 is the population variance of u_i, (3.4), and S_{xe^2}, S_{x^2e} are the population covariances of x_i and e_i^2, x_i^2 and e_i respectively.

(b) $$v_{1r} = \left(\frac{1-f}{n}\right) \frac{n-1}{n-2} s_e^2 - \frac{1-f}{n-2} \frac{\bar{u}^2}{s_x^2} + O_p(n^{-2.5}) .$$

If $f = O(n^{-0.5})$, then

(c) $$\text{MSE}(\hat{\bar{y}}_{1r}) = V + \frac{1}{n^2} \left(2 S_e^2 - \frac{S_u^2}{S_x^2} + \frac{4 S_{x^2e}^2 + 2 S_{xe^2} U_3}{S_x^4}\right) + O(n^{-2.5}) .$$

If $f = O(n^{-0.5})$ is relaxed to $f = o(1)$ in Theorem 4.1, $O(n^{-2.5})$ should be changed to $o(n^{-2})$. The same applies to the results of Section 4.2.

4.2. <u>Bias Behavior of</u> v_{1r}, v_L, v_g, v_H, v_D, and v_J

Throughout this subsection we assume $f = O(n^{-0.5})$. For any variance estimator v, we denote its <u>bias</u> for estimating MSE($\hat{\bar{y}}_{1r}$) by $B(v) = E(v) - MSE(\hat{\bar{y}}_{1r})$. The biases of the six variance estimators are given below:

$$B(v_{1r}) = -\frac{1}{n^2}\left(S_e^2 + \frac{2S_{xe^2}U_3 + 4S^2_{x^2e}}{S_x^4}\right) + O(n^{-2.5}), \qquad (4.1)$$

$$B(v_L) = -\frac{1}{n^2}\left[\frac{2S_{xe^2}U_3 + 4S^2_{x^2e}}{S_x^4}\right] + O(n^{-2.5}), \qquad (4.2)$$

$$B(v_g) = \frac{1}{n^2}\left[-S_e^2 - \frac{2S_{xe^2}U_3 + 4S^2_{x^2e}}{S_x^4} - g\frac{S_{xe^2}}{\bar{X}} + \frac{g(g+1)}{2}\frac{S_x^2S_e^2}{\bar{X}^2}\right] + O(n^{-2.5}), \qquad (4.3)$$

$$B(v_D) = O(n^{-2.5}), \qquad (4.4)$$

$$B(v_H) = -\frac{1}{n^2}\left(S_e^2 + \frac{S_u^2}{S_x^2}\right) + O(n^{-2.5}), \qquad (4.5)$$

$$B(v_J) = \frac{1}{n^2}\frac{S_u^2}{S_x^2} + O(n^{-2.5}). \qquad (4.6)$$

Formula (4.5) follows from (3.16) and (4.4); (4.6) from (3.18) and (4.4). The others are proved in the Appendix.

From (4.1) and (4.2), it is easy to see that if

$$S_{xe^2} U_3 \geq 0 , \tag{4.7}$$

then v_L is less downward biased than v_{1r}. In fact, $S_{xe^2} U_3 \geq 0$ for all six populations studied in Royall and Cumberland(1981). Therefore, as expected from our results, both v_{1r} and v_L underestimate MSE($\hat{\bar{y}}_{1r}$) for these populations. See Table 1.

The leading terms of B(v_g) is a quadratic function in g with positive coefficient for the quadratic term. One can easily check that the minimum of B(v_g) occurs at $g = g_{opt}-0.5$, where g_{opt} is defined in Theorem 2.1. Furthermore, if $S_{xe^2} U_3 \geq 0$, then this minimum corresponds to the largest negative bias of v_g. This observation agrees with the empirical study of the next section.

We next observe that v_D, v_J have biases of the order n^{-2} , whereas v_D has a smaller order bias. Up to the order n^{-2}, v_H underestimates MSE($\hat{\bar{y}}_{1r}$), v_J overestimates MSE($\hat{\bar{y}}_{1r}$) and v_D is unbiased in the sense that its leading term is of order $n^{-2.5}$. For the ratio estimator $\hat{\bar{y}}_R$, the overestimation of v_J for MSE($\hat{\bar{y}}_R$) was proved by Wu(1982b). The above observations are supported by the simulation study in Section 5 (Table 2) and an empirical study on six natural populations with sample size 32 in Royall and Cumberland (1981, p.926), on which the following table is based.

Table 1. Relative Bias $B(v)/MSE(\hat{\bar{y}}_{1r})$ of Five Estimators

Population	v_{1r}	v_L	v_H	v_D	v_J
Cancer	-.14	-.12	-.12	-.06	.09
Cities	-.06	-.04	-.04	-.01	.04
Counties 60	-.15	-.14	-.08	-.02	.16
Counties 70	-.14	-.13	-.16	-.07	.14
Hospitals	-.04	-.03	-.02	.01	.06
Sales	-.24	-.21	-.19	-.12	.11

5. Empirical Study

5.1. Populations Under Study and Simulation Procedure

In Sections 3 and 4, the asymptotic behavior of the variance estimators were studied. One may ask whether these results are applicable to moderate sample size. The variance estimators given in Section 2 will be compared empirically on six natural populations. For a detailed description of these populations, see Royall and Cumberland (1981). The procedure described below was conducted on the UNIVAC 1100 at the University of Wisconsin-Madison. The uniform numbers were generated according to subroutine RANUN.

We draw 1000 simple random samples of size 32 from each population whose size ranges from 125 to 393. For each sample chosen, we compute the regression estimate $\hat{\bar{y}}_{1r}$, sample mean $\bar{x}$ and variance estimators v_0, v_1, v_2, $v_{\hat{g}}$, v_L, v_H, v_D and v_J. For each simulated

sample and each variance estimate v, we also compute the t-statistic

$$t = \frac{\hat{\bar{y}}_{1r} - \bar{Y}}{v^{1/2}}, \tag{5.1}$$

and the $(1 - \alpha)$ confidence interval for $\bar{Y}$

$$(\hat{\bar{y}}_{1r} - t_{\alpha/2}(30)\, v^{1/2}, \hat{\bar{y}}_{1r} + t_{\alpha/2}(30)\, v^{1/2}), \tag{5.2}$$

where $t_{\alpha/2}(30)$ is the upper $\alpha/2$ percentile of the t-distribution with 30 d.f.

The unconditional behavior of the estimators can be studied by taking the average of the corresponding quantity among all 1000 samples. For example, the MSE($\hat{\bar{y}}_{1r}$) is calculated as $1000^{-1}\Sigma(\hat{\bar{y}}_{1r} - \bar{Y})^2$ over the 1000 simulated samples, and the bias of a given variance estimator v is calculated as $1000^{-1}\Sigma v$ - MSE($\hat{\bar{y}}_{1r}$) over the same 1000 samples.

To study their conditional behavior on $\bar{x}$, we divide the 1000 samples into groups according to the following procedure. Rearrange the 1000 samples in increasing order of $\bar{x}$; divide the 1000 samples into 10 groups so that the first group has 100 samples whose $\bar{x}$ values are the smallest, the next group contains samples with the next 100 smallest $\bar{x}$ values, and so on. Within each group, we compute the average of $\bar{x}$, v , and the actual percentage coverage of each associated confidence interval.

The following three criteria will be used to compare the preformance of the variance estimators: their mean squared error (MSE) and

bias, and the coverage probability of the associated confidence interval. The simulation results are summarized in Tables 2 and 3.

5.2. MSE of v

The pattern is similar to that of Wu and Deng(1983) for the ratio estimator.

(a) $v_{\hat{g}}$ has smaller and often the smallest MSE among all the estimators considered. This is consistent with the asymptotic result of Section 3.

(b) Among v_0, v_1 and v_2, the best performer is the one closest to g_{opt}.

(c) The jackknife variance estimator v_J has the largest MSE among all variance estimators considered.

(d) Among v_H, v_D and v_J, v_H has the smallest MSE.

(e) v_H has bigger MSE than v_0, v_1, v_2, $v_{\hat{g}}$ and v_L.

5.3. Bias of v

(a) All estimators under consideration, except v_J , are consistently downward biased. The downward bias of v_H is predicted in (4.5). Since $S_{xe^2} U_3 \geq 0$ for all six populations, the downward bias of v_0 and v_L is predicted in (4.1) and (4.2).

(b) The estimator v_J is always upward biased while v_D does not show any pattern. This is again well predicted in (4.4) and (4.6).

(c) v_D has the smallest absolute bias among all the estimators. The reason is that v_D is the only estimator with a lower order bias.

(d) v_L has a smaller bias than v_0, v_1, v_2, and $v_{\hat{g}}$.

Table 2. Root mean-square error and bias* of v's

	Population					
	1	2	3	4	5	6
v_0	2.91	52.6	13.6	22.0	6.75	24.9
	(-1.2)	(-5.2)	(-10.0)	(-5.6)	(-1.6)	(-13.8)
v_1	2.51	51.3	13.1	19.0	6.12	20.9
	(-1.3)	(-5.8)	(-10.0)	(-6.9)	(-1.8)	(-14.7)
v_2	2.39	54.4	13.3	18.3	6.24	19.4
	(-1.3)	(-4.9)	(- 9.4)	(-7.3)	(-1.7)	(-13.7)
$v_{\hat{g}}$	2.49	51.1	13.2	18.9	6.24	20.7
	(-1.4)	(-6.9)	(- 9.1)	(-7.0)	(-1.8)	(-14.5)
v_L	2.92	55.1	13.1	22.6	6.85	24.2
	(-1.0)	(-1.2)	(- 9.0)	(-4.8)	(-1.1)	(-11.7)
v_H	2.74	59.4	17.9	23.3	7.64	22.0
	(-1.1)	(-1.4)	(- 5.7)	(-5.5)	(-0.9)	(- 9.6)
v_D	3.42	66.1	22.4	30.9	8.92	26.7
	(- .5)	(+ .5)	(- 2.6)	(-1.6)	(+.08)	(- 4.4)
v_J	5.56	84.2	37.1	52.6	11.36	48.1
	(+0.6)	(+16.8)	(+ 5.3)	(+7.1)	(+1.8)	(+ 9.9)
g_{opt}	1.55	1.20	0.88	2.40	1.46	1.53
Unit	1	10000	1000	1000	100	100000

* Bias given inside the parenthesis

5.4. Behavior of the Confidence Intervals

Only the results on populations 1 and 6 are reported in Table 3. They are representative of a bigger study in Deng(1984), which is well summarized by the following conclusions.

(a) Normality of the t-statistic:

(a1) The behavior of the t-statistic is similar to the student t-distribution: the bias and skewness close to zero and standard deviation close to one.

(a2) The t-statistic associated with v_0 has the largest variance while that associated with v_J is the smallest.

(b) Unconditional coverage probability:

(b1) For all six populations, the coverage probability is lower than the nominal level $1 - \alpha$.

(b2) The confidence interval associated with v_J has the closest coverage probability to the nominal level while that associated with v_0 has the lowest coverage probability.

(b3) The confidence interval associated with v_J has the best performance among all estimators considered. The superior performance of v_J can be explained in part by the large values of $E(v_J)$.

(b4) Among v_0, v_1 and v_2, v_2 is the best and v_0 the worst.

(b5) Among v_H, v_D and v_J, v_J is the best and v_H the worst. This may partly be explained by the results in Section 3 where v_H was shown to be stochastically smaller than v_D and v_D smaller than v_J.

(c) Conditional coverage probability:

(c1) We can clearly see the excellent performance of the conditional coverage probabilities associated with v_J. They do not fluctuate very much as $\bar{x}$ varies.

(c2) Compared with the other estimators, the coverage probabilities associated with v_H, v_D, v_J are pretty stable over $\bar{x}$, whereas those associated with v_0, v_L, $v_{\hat{g}}$ are increasing in $\bar{x}$. For example, in population 1, the actual coverage probability of the 95% confidence interval associated with v_0 in the first group is as low as 73% and in the last group as high as 99%.

(c3) Among v_0, v_1, v_2, v_2 has the most stable conditional coverage probabilities.

(c4) Among v_H, v_D, v_J, v_J has bigger coverage probabilities than that of v_D for each group; and v_D bigger than v_H. This again can be explained by our asymptotic results in Section 3.

(c5) For "nearly" balanced samples (i.e. $\bar{x}$ close to $\bar{X}$), all estimators perform similarly. For example, for each population the 5-th and 6-th groups have similar coverage probabilities for all estimators.

Table 3. Coverage probabilities of the t-intervals in (5.2) and descriptive statistics of t in (5.1) based on 1000 samples

Population 1

	99%	95%	90%	Bias	Var.	Skew.	Kurt.
t_0	94.3	88.5	80.5	-.1311	1.9640	-.0925	4.7126
t_1	95.1	89.3	82.5	-.1223	1.7470	-.0593	4.3104
t_2	96.4	89.8	84.2	-.1151	1.6067	-.0203	3.8002
t_g	94.9	89.0	82.4	-.1217	1.8276	-.0352	4.4895
t_L	94.7	89.1	82.0	-.1232	1.8255	-.0760	4.5755
t_H	96.0	90.1	83.5	-.1105	1.6160	-.0225	4.1515
t_D	96.4	91.4	85.2	-.1077	1.4987	-.0581	4.2578
t_J	97.3	92.7	87.6	-.1050	1.3053	-.1010	4.3977

Conditional 95% C.I. coverage probability

$\bar{x}$	t_0	t_1	t_2	t_H	t_D	t_J	t_g	t_L
76.0	73	81	85	87	88	91	80	79
88.4	78	81	84	82	84	89	81	80
96.5	76	77	82	81	82	85	77	76
102.6	84	85	88	88	89	90	84	84
109.2	94	94	95	95	95	96	94	94
115.4	87	86	85	87	91	91	87	87
121.9	96	95	92	95	96	96	96	96
128.2	97	96	95	95	97	97	97	97
137.2	99	97	96	95	96	96	97	99
157.1	99	98	96	96	96	96	97	99

Population 6

	99%	95%	90%	Bias	Var.	Skew.	Kurt.
t_0	91.8	84.1	77.4	-.0793	2.4208	-.1271	5.2355
t_1	94.2	85.7	79.4	-.0784	2.0163	-.1136	4.5610
t_2	95.8	87.9	80.2	-.0772	1.7840	-.0474	4.0702
t_g	93.8	85.1	79.1	-.0753	2.1614	-.1846	5.2800
t_L	93.9	86.1	79.5	-.0886	2.0438	-.2675	5.3736
t_H	96.2	89.4	82.4	-.0744	1.7426	-.2565	5.5570
t_D	96.6	91.2	85.0	-.0458	1.5194	-.1137	5.1693
t_J	98.4	93.9	89.2	-.0195	1.1383	.1069	4.3019

Conditional 95% C.I. coverage probability

$\bar{x}$	t_0	t_1	t_2	t_H	t_D	t_J	t_g	t_L
14.3	60	70	82	88	87	90	63	73
16.7	76	79	87	85	87	91	78	78
18.2	72	74	79	79	81	87	73	73
19.7	77	87	89	90	91	95	84	81
21.2	85	88	92	90	95	98	87	85
22.8	90	90	90	90	93	96	90	90
24.3	92	91	90	90	92	93	92	92
26.5	94	90	87	92	94	95	92	94
29.5	95	91	90	93	95	96	94	95
36.9	100	97	93	97	97	98	98	100

Appendix

Proof of Lemma 3.1. Parts (b) and (c) follow from (a) and formulas (13) and (14) of Wu(1982a). To prove (a), from formula (7.31) of Cochran(1977) and formulas (3.2),(3.5) and (3.6), we have

$$\sum_{i=1}^{n} [(y_i - \bar{y}) - b(x_i - \bar{x})]^2 = \sum_{i=1}^{n} (e_i - \bar{e})^2 - \delta_n^2 \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$= \sum_{i=1}^{n} (e_i - \bar{e})^2 - \frac{n\,\bar{u}^2}{s_x^2} + 0_p(n^{-0.5}) = \sum_{i=1}^{n} e_i^{\,2} + 0_p(1) \ . \qquad (A3.1)$$

This proves part (a).

Proof of Lemma 3.2. Note that

$$\bar{x}_r - \bar{x} = \frac{N\,\bar{X} - n\,\bar{x}}{N - n} - \bar{x} = \frac{-1}{1-f}(\bar{x} - \bar{X}) \ . \qquad (A3.2)$$

From (3.14) and (A3.2), the numerator of α_i in (2.15) is equal to

$$= 1 - 2\frac{p}{1-f}(x_i - \bar{x}) + (\frac{p}{1-f})^2(x_i - \bar{x})^2 + \frac{f}{1-f}$$

$$= \frac{1}{1-f}[1 - 2p(x_i - \bar{x}) + p^2(x_i - \bar{x})^2] + 0_p(n^{-1.5}) \ . \qquad (A3.3)$$

We used the facts $p^2 = 0_p(n^{-1})$ and $f = 0(n^{-0.5})$ in deriving (A3.3). From (2.15) and (A3.3) , we obtain

$$\alpha_i = (1-f)^{-1}(1 - p(x_i - \bar{x}))^2/(1 - q(x_i - \bar{x})^2) + 0_p(n^{-1.5}), \qquad (A3.4)$$

which easily implies (3.12). Formula (3.13) follows from (3.12) and

$$(1 - q(x_i - \bar{x})^2)^{-1} = 1 + q(x_i - \bar{x})^2 + 0_p(n^{-1.5}) \ .$$

Proof of Lemma 3.3.

From $w_i = 0_p(n^{-1})$ and $k_i = 0_p(n^{-1})$, β_i in v_H satisfies

$$\beta_i = r_i^{\,2} + 0_p(n^{-2}).$$

From (2.14), we have

$$v_H = \left(\frac{1-f}{n}\right)^2 \sum_{i=1}^{n} \hat{e}_i^2 \; r_i^2 + f\frac{1-f}{n(n-2)} \sum_{i=1}^{n} \hat{e}_i^2 + O_p(n^{-2.5}) .$$

From (3.14) and (A3.2),

$$v_H = \left(\frac{1-f}{n}\right)^2 \sum_{i=1}^{n} \hat{e}_i^2 \left(1 - \frac{p}{1-f}(x_i - \bar{x})\right)^2 + f(1-f) \frac{1}{n^2} \sum_{i=1}^{n} \hat{e}_i^2 + O_p(n^{-2.5})$$

$$= \frac{1-f}{n^2} \sum_{i=1}^{n} \hat{e}_i^2 - 2 \frac{1-f}{n^2} p \sum_{i=1}^{n} \hat{e}_i^2 (x_i - \bar{x}) + \frac{p^2}{n^2} \sum_{i=1}^{n} \hat{e}_i^2 (x_i - \bar{x})^2 + O_p(n^{-2.5}),$$

which gives the desired result since $p^2 = O_p(n^{-1})$ and $f = O(n^{-0.5})$.

<u>Proof of Lemma 3.5</u>. From formula (6.1) of Royall and Cumberland (1978, p.357), we have

$$\sum_{i=1}^{n} (\hat{\theta}_{(i)} - \hat{\theta}_{(.)})^2 = N^{-2}\left[\sum_{i=1}^{n} (1 + g_i)^2 \hat{e}_i^2 (1 - k_i)^{-2}\right] + r_n , \qquad \text{(A3.5)}$$

where

$$r_n = - \frac{1}{nN^2} \left[\sum_{i=1}^{n} \hat{e}_i(1 - k_i)^{-1} + \sum_{i=1}^{n} g_i \hat{e}_i k_i(1 - k_i)^{-1}\right]^2 , \qquad \text{(A3.6)}$$

$$g_i = \frac{(N-n)}{n}(1, x_i) \begin{bmatrix} 1 & \bar{x} \\ \bar{x} & \bar{x}^{(2)} \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ \bar{x}_r \end{bmatrix}$$

$$= \frac{(N-n)}{n}\left(1 + \frac{(x_i - \bar{x})(\bar{x}_r - \bar{x})}{g(s)}\right)$$

$$= \frac{1-f}{f} - \frac{1}{f} \frac{(x_i - \bar{x})(\bar{x} - \bar{X})}{g(s)} = O_p(n^{0.5}) , \qquad \text{(A3.7)}$$

$\bar{x}^{(2)}$ is the second sample moment of x and k_i is defined in (2.19). We then show $r_n = O_p(n^{-3})$. From $g_i = O_p(n^{0.5})$ and $k_i = O_p(n^{-1})$, the second term inside the square bracket of (A3.6) is $O_p(n^{0.5})$. Its first term

$$\sum_{i=1}^{n} \hat{e}_i(1 - k_i)^{-1} = \sum_{i=1}^{n} \hat{e}_i + \sum_{i=1}^{n} \hat{e}_i k_i(1 - k_i)^{-1}$$

$$= \sum_{i=1}^{n} \hat{e}_i + 0_p(1) = 0_p(n^{0.5}) .$$

Therefore $r_n = 0_p(N^{-2}) = 0_p(f^2 n^{-2}) = 0_p(n^{-3})$. The proof is completed by applying $1 + g_i = f^{-1} (1 - p(x_i - \bar{x}))^2$ and $(1 - k_i) = (\frac{n-1}{n}) (1 - q(x_i - \bar{x})^2)$ to (A3.5).

Proof of Lemma 3.6. It follows easily from Lemma 3.2, (3.11), Lemma 3.5 and

$$(1 - q(x_i - \bar{x})^2)^{-2} = 1 + 2q(x_i - \bar{x})^2 + 0_p(n^{-2}) .$$

Proof of Theorem 4.1. Using (3.3) and (3.7), we can show that

$$(\hat{\bar{y}}_{1r} - \bar{Y})^2 = (\bar{e} - \delta_n(\bar{x} - \bar{X}))^2$$

$$= \bar{e}^2 - 2 \frac{\bar{e}\,\bar{u}(\bar{x} - \bar{X}) - \bar{e}^2(\bar{x} - \bar{X})^2}{s_x^2} + \frac{\bar{u}^2(\bar{x} - \bar{X})^2}{s_x^4}$$

$$+ 2 \frac{\bar{e}\,\bar{u}(\bar{v} - \bar{V})(\bar{x} - \bar{X})}{s_x^4} + 0_p(n^{-2.5}) \tag{A4.1}$$

To compute the expectation of (A4.1), we need the following formulas:

$$E(\bar{e}\,\bar{u}(\bar{x} - \bar{X})) = \frac{(1 - f)(1 - 2f)}{n^2} S_u^2 + 0(n^{-2.5}) \tag{A4.2}$$

$$E(\bar{u}^2(\bar{x} - \bar{X})^2) = (\frac{1-f}{n})^2(S_u^2 S_x^2 + 2(S_{xu})^2) + 0(n^{-2.5}) \tag{A4.3}$$

$$E(\bar{e}^2(\bar{x} - \bar{X})^2) = (\frac{1-f}{n})^2(S_e^2 S_x^2 + (S_{xe})^2) + 0(n^{-2.5}) \tag{A4.4}$$

$$= (\frac{1-f}{n})^2 S_e^2 S_x^2 + 0(n^{-2.5}) \tag{A4.5}$$

$E(\bar{e}\,\bar{u}(\bar{v}-\bar{V})(\bar{x}-\bar{X}))$

$$=\left(\frac{1-f}{n}\right)^2(S_{eu}S_{vx}+S_{ev}S_{xu}+S_{xe}S_{uv})+O(n^{-2.5}) \qquad (A4.6)$$

$$=\left(\frac{1-f}{n}\right)^2(S_{xe^2}U_3+S_{xu}{}^2)+O(n^{-2.5}) \qquad (A4.7)$$

Formulas (A4.3) and (A4.4) follow easily from Sukhatme and Sukhatme (1970, p.192, (9)). Formulas (A4.2) and (A4.6) follow from Theorems 1 and 2 of Nath(1968). Formulas (A4.5) and (A4.7) hold, because $S_{xe}=0$, $S_{vx}=U_3$, $S_{eu}=S_{xe^2}$, $S_{xu}=S_{x^2e}$ and $S_{ev}=S_{xu}$. Combining (A4.1)-(A4.7) we obtain

$$MSE(\hat{\bar{y}}_{1r})=E(\hat{\bar{y}}_{1r}-\bar{Y})^2$$

$$=\left(\frac{1-f}{n}\right)S_e^2-2\frac{1}{n^2}\left[\frac{(1-f)(1-2f)\,S_u^2-(1-f)^2\,S_e^2\,S_x^2}{S_x^2}\right]$$

$$+\left(\frac{1-f}{n}\right)^2\frac{S_u^2\,S_x^2+2(S_{xu})^2}{S_x^4}+O(n^{-2.5})$$

and establish (a). If $f=O(n^{-0.5})$, then part (c) follows easily from (a). Part(b) follows from (A3.1).

<u>Proof of</u> (<u>4</u>.<u>1</u>). By taking expectation of Theorem 4.1 (b), we get

$$E(v_{1r})=\left(\frac{1-f}{n}\right)S_e^2+\frac{1}{n^2}(S_e^2-(1-f)\,S_u^2/S_x^2)+O(n^{-2.5}),$$

which and Theorem 4.1(c) imply the result.

<u>Proof of</u> (<u>4</u>.<u>2</u>). Note that

$$(\bar{x}-\bar{X})^2/\left[\left(\frac{1-f}{n}\right)\sum_{i=1}^{n}(x_i-\bar{x})^2\right]=(\bar{x}-\bar{X})^2/S_x^2+O_p(n^{-1.5}).$$

This implies, using (2.12),

$$v_L = v_{1r} + \left(\frac{1-f}{n}\right)\frac{S_e^2(\bar{x} - \bar{X})^2}{S_x^2} + 0_p(n^{-2.5}) ,$$

$$E(v_L) = E(v_{1r}) + \frac{1}{n^2} S_e^2 + 0(n^{-2.5}) , \qquad \text{(A4.8)}$$

which and (4.1) imply the result.

Proof of (4.3). From

$$\left(\frac{\bar{X}}{\bar{x}}\right)^g = 1 - g\frac{(\bar{x} - \bar{X})}{\bar{X}} + \frac{g(g+1)}{2}\frac{(\bar{x} - \bar{X})^2}{\bar{X}^2} , + 0_p(n^{-1.5})$$

$$v_g = v_{1r} + [(\frac{1-f}{n}) s_e^2 + 0_p(n^{-2})]$$

$$[- g(\bar{x} - \bar{X})/\bar{X} + g(g+1)/2(\bar{x} - \bar{X})^2/\bar{X}^2 + 0_p(n^{-1.5})]$$

$$= v_{1r} + (\frac{1-f}{n})[- g\, s_e^2\frac{(\bar{x} - \bar{X})}{\bar{X}} + \frac{g(g+1)}{2} s_e^2\frac{(\bar{x} - \bar{X})^2}{\bar{X}^2}] + 0_p(n^{-2.5})$$

$$= v_{1r} + (\frac{1-f}{n})[- g[S_e^2 + (s_e^2 - S_e^2)]\frac{(\bar{x} - \bar{X})}{\bar{X}}$$

$$+ \frac{g(g+1)}{2} S_e^2\frac{(\bar{x} - \bar{X})^2}{\bar{X}^2}] + 0_p(n^{-2.5}) .$$

Taking the expectation, we get

$$E(v_g) = E(v_{1r}) + (\frac{1-f}{n})^2[- g\frac{S_{xe^2}}{\bar{X}} + \frac{g(g+1)}{2}\frac{S_e^2 S_x^2}{\bar{X}^2}] + 0(n^{-2.5}) ,$$

which together with Theorem 4.1(c) gives the result.

To prove (4.4), we need the following formulas and Lemmas A4.1 and A4.2. Formulas (A4.9)-(A4.11) find the leading terms of p, p^2 and q, defined in Lemma 3.2,

$$p = \frac{(\bar{x} - \bar{X})}{S_x^2} - \frac{(\bar{x} - \bar{X})(\bar{v} - \bar{V})}{S_x^4} + O_p(n^{-1.5}) , \qquad \text{(A4.9)}$$

$$p^2 = \frac{(\bar{x} - \bar{X})^2}{S_x^4} + O_p(n^{-1.5}) , \qquad \text{(A4.10)}$$

$$q = \frac{1}{n\, S_x^2} + O_p(n^{-1.5}) , \qquad \text{(A4.11)}$$

where v_i and $\bar{V}$ are defined in (3.4).

<u>Lemma</u> <u>A4</u>.<u>1</u>.

$$\sum_{i=1}^{n} \hat{e}_i^{\,2} (x_i - \bar{x})^2 = n\, S_u^2 + O_p(n^{0.5}) , \qquad \text{(A4.12)}$$

where

$$S_u^2 = \frac{1}{N-1} \sum_{i=1}^{N} u_i^{\,2} , \quad u_i = e_i (x_i - \bar{X}) .$$

<u>Proof</u>. From $\delta_n = b - B = O_p(n^{-0.5})$,

$$\hat{e}_i = (e_i - \bar{e}) - \delta_n (x_i - \bar{x}) = e_i + O_p(n^{-0.5}) \qquad \text{(A4.13)}$$

and $\hat{e}_i^{\,2} = e_i^{\,2} + O_p(n^{-0.5})$, which and $(x_i - \bar{x})^2 = (x_i - \bar{X})^2 + O_p(n^{-0.5})$ imply

$$\hat{e}_i^{\,2} (x_i - \bar{x})^2 = e_i^{\,2} (x_i - \bar{X})^2 + O_p(n^{-0.5}) , \qquad \text{(A4.14)}$$

from which the result follows easily.

<u>Lemma</u> <u>A4</u>.<u>2</u>.

$$\sum_{i=1}^{n} \hat{e}_i^{\,2} (x_i - \bar{x}) = n[\, \bar{w} - (\bar{x} - \bar{X})\, S_e^2 - 2\, \bar{u} \frac{S_{xu}}{S_x^2}] + O_p(1), \qquad \text{(A4.15)}$$

where

$$\bar{w} = \frac{1}{n} \sum_{i=1}^{n} w_i , \quad w_i = e_i^{\,2} (x_i - \bar{X}) .$$

Proof. From (A4.13), we have

$$\hat{e}_i^2 = e_i^2 - 2\delta_n e_i(x_i - \bar{x}) - 2 e_i \bar{e} + O_p(n^{-1})$$

$$= e_i^2 - 2\delta_n e_i(x_i - \bar{X}) - 2 e_i \bar{e} + O_p(n^{-1}) .$$

Ignoring terms of order n^{-1}, we find

$$\hat{e}_i^2 (x_i - \bar{x}) = \hat{e}_i^2 [(x_i - \bar{X}) - (\bar{x} - \bar{X})]$$

$$= e_i^2(x_i - \bar{X}) - 2 \delta_n e_i(x_i - \bar{X})^2 - 2 e_i \bar{e}(x_i - \bar{X})$$

$$- e_i^2(\bar{x} - \bar{X}) + O_p(n^{-1})$$

which implies Lemma A4.2, by using (3.2) and (3.5).

Proof of (4.4). Using (A4.9) and Lemma A4.2, the second term of v_D in (3.13) is

$$- 2p \sum_{i=1}^{n} \hat{e}_i^2 (x_i - \bar{x})$$

$$= - 2 \left[\frac{(\bar{x} - \bar{X})}{s_x^2} - \frac{(\bar{x} - \bar{X})(\bar{v} - \bar{V})}{s_x^4} + O_p(n^{-1.5})\right] .$$

$$\left[n (\bar{w} - (\bar{x} - \bar{X}) s_e^2 - 2 \bar{u}\frac{s_{xu}}{s_x^2}) + O_p(1)\right]$$

$$= - 2 n \left[\frac{\bar{w}(\bar{x} - \bar{X})}{s_x^2} - \frac{\bar{W}(\bar{x} - \bar{X})(\bar{v} - \bar{V})}{s_x^4}\right.$$

$$\left. - \frac{(\bar{x} - \bar{X})^2 s_e^2}{s_x^2} - 2 \bar{u}(\bar{x} - \bar{X})\frac{s_{xu}}{s_x^4}\right] + O_p(n^{-0.5}) . \qquad \text{(A4.16)}$$

The third term of v_D in (3.13) can be simplified by using (A4.10),(A4.11) and Lemma A4.1,

$$(p^2 + q) \sum_{i=1}^{n} \hat{e}_i^{\,2} (x_i - \bar{x})^2$$

$$= (\frac{(\bar{x} - \bar{X})^2}{S_x^4} + \frac{1}{n\, S_x^2} + O_p(n^{-1.5}))(n\, S_u^2 + O_p(n^{0.5}))$$

$$= n(\bar{x} - \bar{X})^2 \frac{S_u^2}{S_x^4} + \frac{S_u^2}{S_x^2} + O_p(n^{-0.5}) \;. \qquad \text{(A4.17)}$$

Combining (A4.16), (A4.17) and Lemma 3.2, we get

$$v_D = (1 - \frac{1}{n-1}) v_{1r} - 2 \frac{1}{n} [\frac{\bar{w}(\bar{x} - \bar{X})}{S_x^2} - \frac{\bar{W}(\bar{x} - \bar{X})(\bar{v} - \bar{V})}{S_x^4}$$

$$- \frac{(\bar{x} - \bar{X})^2 S_e^2}{S_x^2} - 2\, \bar{u}(\bar{x} - \bar{X})\frac{S_{xu}}{S_x^4}]$$

$$+ \frac{1}{n} (\bar{x} - \bar{X})^2 \frac{S_u^2}{S_x^4} + \frac{1}{n^2} \frac{S_u^2}{S_x^2} + O_p(n^{-2.5}) \;. \qquad \text{(A4.18)}$$

Collecting the leading terms of $E(v_D)$, we have

$$E(v_D) = E(v_{1r}) - \frac{1}{n} E(v_{1r}) + \frac{1}{n^2} [S_x^2 \frac{S_u^2}{S_x^4} + \frac{S_u^2}{S_x^2}]$$

$$- 2 \frac{1}{n^2} [\frac{S_u^2}{S_x^2} - \frac{S_{xe^2} U_3}{S_x^4} - \frac{S_x^2 S_e^2}{S_x^2} - 2 \frac{S_{xu}^{\,2}}{S_x^4}] + O(n^{-2.5})$$

$$= E(v_{1r}) + \frac{1}{n^2} [S_e^2 + 2 \frac{S_{xe^2} U_3}{S_x^4} + 4 \frac{S_{xu}^{\,2}}{S_x^4}] + O(n^{-2.5}) \;. \qquad \text{(A4.19)}$$

In writing (A4.19), we used $f = O(n^{-0.5})$ and $E(v_{1r}) = n^{-1} S_e^2 + O(n^{-1.5})$. The result follows from (A4.19) and Theorem 4.1(c).

REFERENCES

Cochran, W. G. (1977), Sampling Techniques, 3rd edition. New York: Wiley

Deng, L. Y. (1984), Statistical Inference in Finite Population Sampling When Auxiliary Information is Available. Ph.D. Thesis, University of Wisconsin, Madison

Fuller, W. A., (1977), "A note on regression estimation for sample surveys," unpublished manuscript.

Fuller, W. A. (1981), "Comment on a paper by Royall and Cumberland," Journal of the American Statistical Association, 76, 78-80.

Iglehart, D. L. (1978), "The regenerative method for simulation analysis," in Current Trends in Programming Methology III, eds. K. M. Chandy and R. T. Yeh. Englewood Cliffs: Prentice-Hall, 57-71.

Nath, S. N. (1968), "On product moments from a finite universe," Journal of the American Statistical Association, 63, 535-541.

Rao, J. N. K. (1968), "Some small sample results in ratio and regression estimation," Journal of the Indian Statistical Association, 6, 160-168.

Rao, J. N. K. (1969), "Ratio and regression estimators," in New Developments in Survey Sampling, ed. N. L. Johnson and H. Smith. New York: Wiley, 213-234.

Rao, J. N. K. (1985), "Ratio estimators," to appear in Encyclopedia of Statistical Sciences Vol. V, eds. S. Kotz and N. L. Johnson. New York: Wiley

Royall, R. M. (1970), "On finite population sampling theory under certain linear regression models ," Biometrika, 57, 377-387.

Royall, R. M. and Cumberland, W. G. (1978), "Variance estimation in finite population sampling," Journal of the American Statistical Association, 73, 351-358.

Royall, R. M. and Cumberland, W. G. (1981), "The finite population linear regression estimator and estimator of its variance - An empirical study ," Journal of the American Statistical Association, 76, 924-930.

Royall, R. M. and Eberhardt, K. R. (1975), "Variance estimates for the ratio estimator ," Sankhya C, 37, 43-52.

Scott, A. J. and Smith, T. M. F. (1969), "Estimation in multi-stage surveys ," Journal of the American Statistical Association, 64, 830-840.

Sukhatme, P. V. and Sukhatme, B. V. (1970), Sampling Theory of Surveys with Application. Ames: Iowa State University Press

Wu, C. F. (1982a), "Estimation of variance of the ratio estimator," Biometrika, 69, 183-189.

Wu, C. F., (1982b), Personal Communication.

Wu, C. F. and Deng, L. Y. (1983), "Estimation of variance of the ratio estimator: An empirical study," in Scientific Inference, Data Analysis, and Robustness, ed. G.E.P. Box, et al. New York: Academic Press, 245-277.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER #2758	2. GOVT ACCESSION NO. AD A149407	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Estimation of Variance of the Regression Estimator		5. TYPE OF REPORT & PERIOD COVERED Summary Report - no specific reporting period
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Lih-Yuan Deng and C. F. Jeff Wu		8. CONTRACT OR GRANT NUMBER(s) MCS-8300140 DAAG29-80-C-0041
9. PERFORMING ORGANIZATION NAME AND ADDRESS Mathematics Research Center, University of Wisconsin 610 Walnut Street Madison, Wisconsin 53706		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS Work Unit Number 4 - Statistics & Probability
11. CONTROLLING OFFICE NAME AND ADDRESS See Item 18 below		12. REPORT DATE October 1984
		13. NUMBER OF PAGES 33
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

National Science Foundation
Washington, DC 20550

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Variance estimator; Design-based; Model-based; Jackknife; Conditional coverage probability

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

For estimating the variance of the regression estimator in simple random sampling without replacement, several design-based and model-based estimators and a new class of estimators are compared. Their second order expressions and biases are derived and compared. Empirical results on the biases and MSE's of the variance estimators and the conditional and unconditional coverage probabilities of their associated t-intervals lend support to the theoretical results and suggest further questions.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

END

FILMED

2-85

DTIC